*Language and Speech, **41**, 1-20 (1998).*

Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time.

ANNETTE DOWD[1], JOHN SMITH and JOE WOLFE[2]
School of Physics
The University of New South Wales
Sydney 2052 AUSTRALIA

Abstract

An acoustic impedance spectrometer was used to measure the frequencies R1 and R2 of the first two resonances of the vocal tract. The measurement was made just outside the mouth, in parallel with the free field, using a new technique that provides precise information about the acoustic response of the vocal tract in real time. Values measured for native speakers for a particular vowel were used as target parameters for subjects who used a visual display of an impedance spectrum of their own vocal tracts as real-time feedback to realise the vocal tract configuration required to pronounce the target vowel. We report the values (R1,R2) for eleven non-nasalised vowels of French. These values are similar to the formant frequencies measured previously for these vowels, and their relative positions in the (R2,R1) plane are similar to those of the same vowels in the (F2,F1) formant plane. The confusion and correct identification of these vowels are shown to be strongly related to their separation in the (R2,R1) plane. We report the results of attempts to imitate six of these vowels by monolingual anglophone subjects. One group used a traditional method of learning pronunciation: they heard the vowel sounds and then attempted to imitate them. Another group also heard the sounds, but were assisted by the vocal tract feedback described above when imitating the target sounds. The acoustic properties and recognizability of the vowels were significantly superior when the subjects used vocal-tract feedback.

Key words

Acoustic impedance, vocal tract, formants, speech trainer, French vowels.

INTRODUCTION

It is widely observed and regretted that adults who learn foreign languages rarely acquire an authentic pronunciation. This observation is usually attributed, at least in part, to the processes of categorisation and interference: adults hear a foreign phoneme and tend to interpret it in terms of a similar phoneme in their native language (e.g. Landercy and Renard 1977; Clark and Yallop 1990). When asked to imitate the foreign phoneme, they usually produce a phoneme similar to one in their native language. In some cases the foreign language may make use of a distinction that does not exist in the native language of the student, and in such cases the student may find it difficult to hear and to reproduce the difference. For example, English makes no use of the difference between /ʊ/ and /y/. Consequently, many native English speakers have difficulty hearing and reproducing the difference between these two phonemes, as for example in the French words 'dessous' [d(ə)su], meaning 'below', and 'dessus' [d(ə)sy] meaning 'above'. The normal mode for learning, both for native and foreign languages, uses primarily auditory feedback: infants and students hear speech sounds and attempt to reproduce some aspects of

---

[1] Current address: Department of Electronic Materials Engineering, Research School of Physical Sciences and Engineering, Australian National University, Canberra ACT 0200 Australia.

[2] To whom correspondence should be addressed. J.Wolfe@unsw.edu.au

those sounds. For adults, however, the problems of categorisation and interference may be regarded as a limitation of this method.

Feedback that does not involve hearing should not suffer from the categorisation problem. We have recently developed a method for measuring the acoustic impedance and other transfer functions of the vocal tract in 'real time' - i.e. several times per second (Wolfe, Smith, Brielbeck and Stocker 1994, 1995; Dowd, Smith and Wolfe 1996). This allows us to determine the acoustic resonances of the tract. In this study we display the frequencies of the first two resonances to provide feedback about vowel pronunciation which is not subject to categorisation and interference. We know of no previous attempt to teach the pronunciation of foreign vowel sounds using visual feedback from direct, non-invasive measurements of the acoustic properties of the tract.

The resonances of the vocal tract are sometimes called formants in acoustical phonetics. 'Formant' is also used to describe a broad frequency band of emitted power. Our technique allows the two quantities to be measured independently, so we are careful to distinguish between the two. We reserve 'formant' for the broad frequency band of emitted power and the terms F1, F2 etc for the frequencies of the formants. We use the terms R1, R2 etc to denote the frequencies of the resonances of the vocal tract. In the source-filter model of voiced speech (Fant 1973) the formants F1 (where present) and F2 are expected to have frequencies close to those of R1 and R2, the closeness being determined by the fundamental frequency F0 of the phonation. This is shown in Figure 1, in which (**a**) is an idealised spectrum of the input from the vocal folds, (**b**) is the gain of the vocal tract in a particular configuration showing two maxima at the resonances R1 and R2, and (**c**) is the spectrum of the output sound. For a low-pitched voice (F0 = 100 Hz in the diagram at left) the peaks in the spectral envelope of the output sound (the formants) are reasonably clear. For a high-pitched voice (F0 = 300 Hz in the diagram at right), it is more difficult to determine the formants accurately.

In principle, the values of formant frequencies may be used as feedback to teach pronunciation of vowel sounds. One difficulty in doing so in practice is that it is difficult to determine the formants precisely, rapidly and automatically. It is possible to extract formants from the speech sound using automated techniques such as linear prediction (Makhoul 1975). There is, however, an inherent limit to the capacity of such methods to yield information about the configuration of the vocal tract because the frequency components present during voiced speech are multiples of the fundamental frequency F0. Even for a low pitched voice (F0 say 100 Hz) this limits the resolution in frequency of properties of the vocal tract. For a high pitched voice (F0 several times higher), the resolution in frequency is correspondingly reduced, and it is sometimes difficult, even for a trained observer, to identify formants accurately from a spectrogram of a woman's or child's voice (see Figure 1). Ideally, the information provided as feedback in speech training should be automated and precise. This can be achieved by using impedance spectroscopy which uses an external broadband sound source to provide acoustic information about the vocal tract with greater frequency resolution. Resonances can be determined rather more accurately using external excitation than formants can be determined using linear prediction (Epps, Smith and Wolfe, 1997).

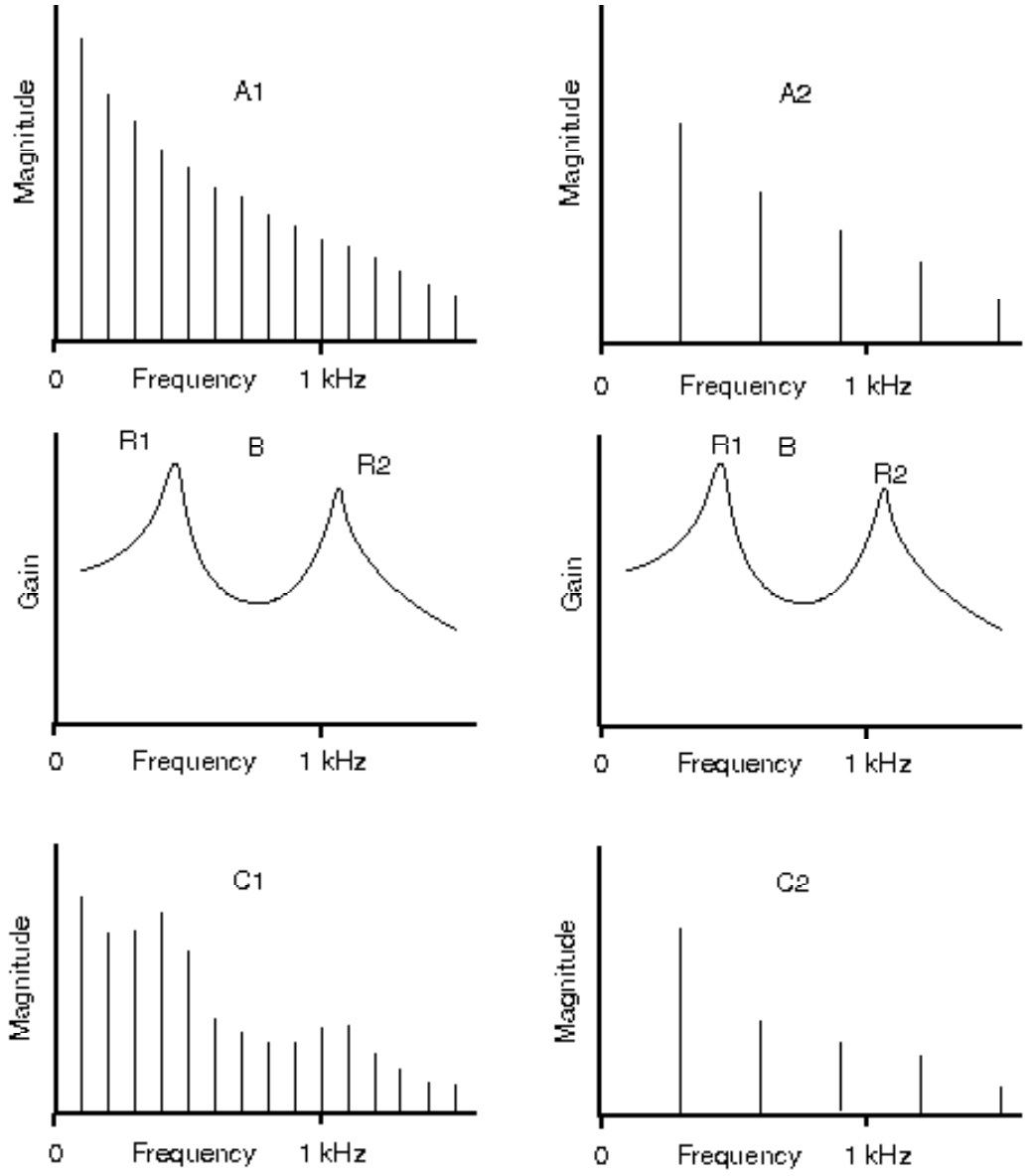

FIGURE 1. A representation of the source-filter model (Fant, 1973) for voiced speech. (a) an idealised spectrum of the input from the vocal folds; (b) the gain of the vocal tract in a particular configuration showing two resonances R1 and R2; (c) the spectrum of the output sound. The series at left shows a low pitched voice (F0 = 100 Hz); that at right shows a high pitched voice (F0 = 300 Hz).

This study concerns only vowels. This choice was made for three reasons. Vowels are sustained sounds, and thus feedback adjustment may be performed continuously in real time. Vowel sounds are produced with the mouth open, so it is technically easier to measure the resonances of the vocal tract using our method. Finally, categorisation and interference are important in limiting the learning of accurate pronunciation of vowels and so feedback on vowel pronunciation is a useful goal for automated speech trainers.

For the target language, we sought a language without lexical tone but with a relatively large number of pure vowels so that fine discrimination could be tested. We were further constrained by the need to find a group of volunteer native speakers who could all produce something approaching a standard pronunciation. French most closely satisfied these criteria.

Rey and Rey-Debove (1985) attribute sixteen[3] vowels to standard spoken French, although the language may be intelligibly spoken with fewer (Liénard 1977). Landercy and Renard (1977) give a reduced list of ten vowels[4] in a necessary and sufficient system for the comprehension of contemporary French. The native speakers participating in the present study claimed to use between thirteen and sixteen.

In this study we measured the first two vocal tract resonances of a sample of native French speakers for the non-nasalised vowels of that language. We also recorded the sounds of those vowels pronounced by those speakers. We then used a subset of our vowel measurements on native French speakers as targets for an imitation study. We did not use all of the vowels in the imitation study because these sessions involved training and we wished to keep their duration less than one hour. We chose six vowels as targets for the feedback study: /e/, /ɛ/, /a/, /ɑ/, /u/ and /y/. These may be considered as pairs that are sometimes confused by non-native speakers of French. The pair /a/ and /ɑ/ are rather similar acoustically and are sometimes confused even by native speakers. The pair /e/ and /ɛ/ are somewhat similar acoustically and Landercy and Renard (1977) do not require their distinction in the reduced list of vowels. On the other hand /u/ and /y/ are quite different acoustically, are never confused by native speakers, but are nevertheless relatively often confused by English speakers, both in recognition and production.

We then used the first two frequencies of the vocal tract resonances and/or the sounds of the vowels as targets for two groups of native English speakers with little experience in speaking foreign languages. One group had a one hour training session in the use of what we call the vocal tract feedback technique, during which they did not hear target sounds (although they could hear their own voices). Members of this group saw, displayed on a computer monitor, the acoustic response of their own vocal tracts, upon which were superimposed the values of R1 and R2 of a speaker whose tract we had previously measured. They were instructed to attempt to match the resonances of their vocal tracts to those of the target speaker. Members of this group then returned later for a session in which they imitated the six target French vowel sounds using both auditory and vocal tract feedback. The second group of subjects imitated the target vowel sounds using conventional, auditory feedback at their first session: members of this group heard recordings of the target sounds spoken by the native speakers and attempted to make sounds similar to these. This second group also returned for a second session. In their second session they attempted to imitate the target using both auditory feedback and vocal tract feedback as described above. In this study we compare the performance of speakers imitating vowel sounds with these different feedback combinations.

In this study, we use only the data for the first two resonances. The first three resonances, especially the first two, are considered important in pronunciation of vowels (Lindblom and Sundberg 1971). The higher resonances are important for voice recognition and speaking or singing style (Sundberg 1987), but have rather less importance in phonology. The restriction to two resonances in this study made training sessions relatively simple. It is quite simple to learn to vary R1 and R2 because the former depends primarily on the jaw opening and latter depends primarily on the tongue position. It is more difficult to vary the frequency of higher resonances

---

[3] /i/ as in 'il'; /e/ as in 'blé'; /ɛ/ as in 'lait'; /a/ as in 'plat'; /ɑ/ as in 'bas'; /ɔ/ as in 'mort'; /o/ as in 'mot'; /u/ as in 'roue'; /y/ as in 'rue'; /ø/ as in 'peu'; /œ/ as in 'peur'; /ə/ as in 'le'; /ɛ̃/ as in 'plein'; /ɔ̃/ as in 'bon'; /ɑ̃/ as in 'sans'; /œ̃/ as in 'brun'.

[4] In this reduced system, no distinction is required between members paired in parentheses in the following list: /i/; (/e/ - /ɛ/); (/a/ - /a/); (/ɔ/ - /o/); /u/; /y/ ; (/ø/ - /œ/); (/ɛ̃/ - /œ̃/; /ɔ̃/; /ɑ̃/. The neutral vowel /ə/ is not cited. When the native French speakers who volunteered for this study were shown the list of sixteen vowels, some of them volunteered, without prompting, that they either used or distinguished fewer than 12 non-nasalised vowels. Some speakers mentioned little difference between /a/ and /ɑ/, and their demonstration of the difference included a longer duration for the former. A few mentioned a small difference between /e/ and /ɛ/, and between /ø/ and /œ/. Two of the native speakers in the target group could be said to be professional speakers, in that they are employed in the French department of the University of New South Wales. These speakers distinguished clearly all 12 non-nasalised vowels listed by Rey and Rey-Debove (1985).

and to explain how it is done. It is also easier for a subject to control two variables simultaneously than three.

MATERIALS AND METHODS

Measurements of acoustic resonances.

A source of carefully synthesized acoustic current was placed near the subject's mouth and the sound pressure recorded by a microphone (see Figure 2). The resonances were then determined from the way in which the subject's vocal tract responded to the synthesized current. The impedance spectrometer and its calibration procedure are described in detail elsewhere (Wolfe et al 1994, 1995). A brief account is given here in Appendix A.

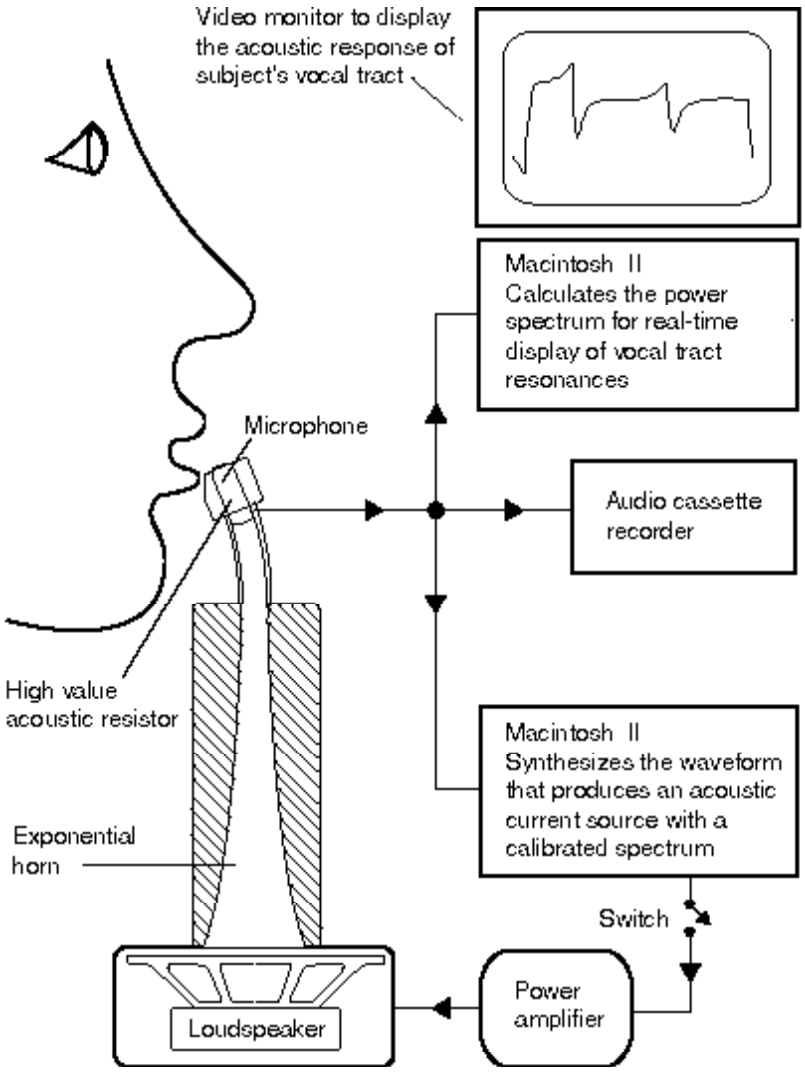

FIGURE 2. The arrangement used to measure the acoustic response of each subject's vocal tract and to display the result to provide vocal tract feedback.

Native speakers.

A group of eight native French speakers, aged between 19 and 64 years, all of whom had come to Australia as adults, were measured separately. They were taught how to raise the velum (soft palate) while not speaking, using a mirror and sometimes feedback from the spectrometer. A

small microphone (8 mm diameter) and the current source (8 mm diameter) of the spectrometer were mounted in a nylon block which lay lightly against the subject's lower lip (Figure 2). Subjects were asked to pronounce in isolation the sixteen French vowels which were presented to them in written form thus: /i/ as in 'pie'; /e/ as in 'thé'; /ɛ/ as in 'paix'; /a/ as in 'patte'; /ɑ/ as in 'pâte'; /ɔ/ as in 'pote'; /o/ as in 'pot'; /u/ as in 'poux'; /y/ as in 'pu'; /ø/ as in 'peu'; /œ/ as in 'peur'; /ə/ as in 'te'; /ɛ̃/ as in 'pin'; /ɔ̃/ as in 'pont'; /ɑ̃/ as in 'pan'; /œ̃/ as in 'un'. These sounds were recorded on an audio cassette tape using the microphone from the spectrometer. The resonances for each vowel were recorded using the following procedure. The native speakers were asked to pronounce and briefly to sustain each vowel, to hold the vocal tract in that position while not phonating, and then to pronounce the vowel again. The response spectrum of the vocal tract was measured. They then phonated again to check that the vocal tract had not changed. This was repeated twice, giving a total of three sets of measurements for each vowel.

The frequencies R1 and R2 of the first two vocal tract resonances were recorded for each vowel for each native speaker. Only one of the volunteer native speakers was male. We recorded his resonances for comparison, but omitted them from further analysis. For the female native speakers, the mean values of each resonance of each vowel were calculated and these mean values were used as the target values for the (female) imitation subjects. For each vowel spoken, we chose as the auditory target a recording of that vowel as pronounced by the speaker whose vocal tract resonances for that vowel were closest to the mean values of the women native speakers. These recordings were used as target sounds for the imitation subjects. The target sounds were thus produced by vocal tracts with resonances R1 and R2 slightly different from the mean values for the whole sample that were used as the targets for vocal tract feedback. The word 'closest' used above refers to displacement in the vowel plane. We quantify this using a deviation-weighted displacement in the plane (R1,R2) of the two resonant frequencies. Over all vowels, the standard deviation σ_2 in $\overline{R2}$ is greater than σ_1 in $\overline{R1}$, so we define the scaled dimensionless displacement between two data ($R1_a$,$R2_a$) and ($R1_b$,$R2_b$) thus:

$$d = \sqrt{\left(\frac{R1_b - R1_a}{\sigma_1}\right)^2 + \left(\frac{R2_b - R2_a}{\sigma_2}\right)^2} \qquad (1)$$

(Dowd, Wolfe and Smith 1996). Here and hereafter 'close' is used to mean having a small displacement d. d may be as large as about 3 or 4 for vocal tracts formed to pronounce very dissimilar vowels, and is zero for identical vocal tract configurations.

Because we had been able to obtain only one male native speaker of French in our target group, we restricted the second part of the study to anglophone women who were asked to imitate the sounds and features of the impedance spectra of the tracts of the native speakers.

Subjects.

A group of eleven adult, female, monolingual native English speakers, aged 21 to 50 years, volunteered to participate in this study. They were not paid for their participation. All were physics students or academic staff at the University of New South Wales. All of them described themselves as having little experience with foreign languages. All had parents who spoke English as their native language. Only two of them had ever lived outside Australia: one 23 year old had lived in New Zealand for the first 21 years of her life, and spoke with a mild New Zealand accent. Another, 24 years old, had lived 11 years in New Zealand, but spoke with no noticeable New Zealand accent. Each of the subjects came to the laboratory for two sessions separated by no more than nine days.

One group of five subjects received training in the vocal tract feedback technique on the first visit, and did not listen to any target sounds. The training consisted in a brief explanation of how to adjust the two resonances shown in the response spectrum of their own vocal tract, and then practice in matching the resonances to given values superimposed upon the screen. The

other group of six received only auditory feedback on their first visit. On their second visit, members of both groups received feedback of both types: auditory and visual.

In all cases the subjects were told that this was a project which aimed to test a new type of acoustic feedback to teach pronunciation. They were told that they would be asked to produce some vowel sounds using one or more methods of feedback. They were not told that the sounds were from French.

Auditory feedback.

For the auditory feedback, subjects listened to vowel sounds via headphones. They listened to a tape recording of each of the six target sounds (/e/, /ɛ/, /a/, /ɑ/, /ɯ/ and /y/) one by one and, in each case, were asked to imitate the sound. They could hear their own versions of the vowel as they attempted to match the target. They were allowed as many repeats as they liked and, when they were satisfied with their match, their version of the sound was recorded three times. Just before the third repetition, a spectrometer measurement of their vocal tract was made just as the velum was involuntarily lifted to begin phonation.

Vocal tract feedback.

For the vocal tract feedback, subjects were taught how to raise the velum (soft palate) while not speaking (the procedure described above for the native speakers). Measurements were made in the same way. Each subject saw on a computer monitor a real time spectrum that displayed the ratio of the acoustic impedance measured at her lower lip with mouth open to that measured with the mouth closed. On this spectrum, each resonance of the tract is usually very obvious as a local maximum followed by a sharp fall and rise (see Figure 3). Simultaneously, each subject saw, superimposed upon the screen, two vertical lines representing the resonance frequencies of the target. They were asked to match the peaks in their spectrum to those of the target, and were told that jaw opening primarily affected R1, and that the position of the tongue (forwards-backwards) primarily affected R2. They were invited to practise the technique with the (R1,R2) values of the first target vowel. This took up to fifteen minutes. The second and successive sounds took less time to learn to imitate (usually less than five minutes). The vowel /y/ usually took longer than the others.

Figure 3 shows the measured spectrum for one of the native speakers for the vowel /œ/. The maxima near 0.5 and 1.5 kHz are due to R1 and R2 respectively. The impedance of the tract is in parallel with that of the laboratory field, which is largely imaginary for the frequencies measured here. The imaginary component of the impedance of the tract changes sign abruptly near each resonance. This leads to the shapes shown here, which have a maximum at the resonant frequency of the radiation loaded vocal tract (Epps, Smith and Wolfe, 1997). The spectrometer outputs signals in the range 0.2 to 2.5 kHz, so any signal outside this range arises from background noise.

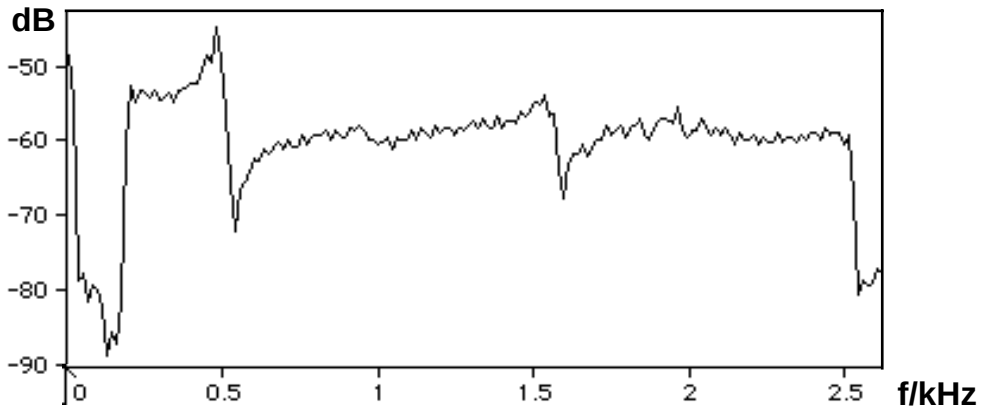

FIGURE 3. The acoustic response of the vocal tract measured for one of the native speakers for the vowel /œ/ .

Acoustic plus vocal tract feedback.

For the combination of the two feedback methods, subjects were presented with the real time response of their own vocal tracts, upon which were superimposed the target values of R1 and R2. They then listened to a tape recording of the target vowel. They were asked to imitate the sound and then to adjust their vocal tracts to match the target. When they were satisfied with their match to the target, they phonated and threw a switch which turned off the acoustic current and turned on a tape recorder using the same microphone (Figure 2). Their spectrum was recorded just before the switch was thrown. The subjects could hear the sound of their own voices when they phonated. The next target was then presented until all six had been recorded.

Recordings for the listening panel.

Each native speaker recorded three repetitions of each of the twelve non-nasalised vowels of French. Each subject in each treatment recorded three repetitions of each of the six non-nasalised vowels chosen for this study. In all cases, the three repetitions were kept together as a block. The blocks were then recorded onto a tape in a randomised sequence for the listening panel. The blocks were not grouped according to vowel, to treatment, to speaker nor to native language. On this tape, an Australian voice announced 'Sound number <n>' which was followed by a block of the three repetitions of the sound. The same Australian voice then announced 'Sound number <n+1>' etc. The tape had a duration of 45 minutes, and multiple copies were made.

Listening panel.

Twelve adult native speakers of French listened to recordings of the sounds of the native speakers and the various imitations. The panel members were unpaid volunteers who were selected for their native language and their willingness to perform this rather tedious task. Their ages ranged from 19 years to 64 years, and their time of residence in Sydney ranged from 2 weeks to 46 years. Four were French citizens temporarily in Sydney to attend university, three were adult immigrants from France who worked as teachers of French, two were French citizens on vacation and three were French citizens temporarily in Sydney for work or family reasons. Members of the listening panel were given a cassette copy of the tape and a printed form. They were asked to read the instructions, to listen to the tape on a hifi player or a personal tape player in quiet conditions, and to mark the form at the same time. They then returned the tape and the form to the laboratory. The form contained instructions and a line for a reply regarding each sound. Each line had the number of the sound followed by 12 words containing the 12 non-nasalised vowels: *pie, thé, paix, patte, pâte, pote, pot, poux, pu, peu, peur* and *te*. The instructions asked the panel members to indicate the word whose vowel was most closely resembled by the sound on the tape. They were asked to listen to the tape just once, without replays.

RESULTS AND DISCUSSION

Resonances of native speakers.

Figure 4 shows the vocal tract resonance data $(\overline{R2},\overline{R1})$ for the seven female native speakers for eleven non-nasalised vowels of French. We stress that these are not measurements of formants. The resonances are thought to be responsible for the formants and so considerable similarities are expected, but the frequency values need not be the same (see Figure 1). In Figure 4, the centre of each ellipse gives the mean values $(\overline{R2},\overline{R1})$ and the slope of the major axis is the correlation between R2 and R1 for that vowel. The lengths of the semiaxes are the standard deviations calculated in those directions. This representation is a simple and objective representation of the data in a way that preserves several features of their spatial distribution.

The results for the first resonance of /i/ were not reliable because the reduced mouth opening provided poor acoustic coupling of the source to the tract for this vowel[5]. These have therefore been omitted. For the nasalised vowels, only one resonance was measured in the range reported here. This result was expected because for the nasalised vowels, the velum is in its lowered position and so the resonance associated with the full length of the vocal tract has less effect on the impedance measured near the mouth. The measured values of (R2,R1) for a male native French speaker (data not shown) were similar to those in Figure 4, but displaced towards the origin. A similar systematic difference in resonance frequencies has been reported by Pham (1995) (see also Högbert 1995).

The relative positions of the vowels in Figure 4 are very similar to those in the articulatory vowel plane presented by, e.g., Landercy and Renard (1977). The vowels occupy much of an approximately triangular section in the (R2,R1) plane. Two of the vertices are at about (800,350) near /u/ and (1500,900) near /a/ (values in Hz). The missing vowel /i/ would form the third vertex. There is also an unoccupied region in the (R2,R1) plane, a roughly triangular shape with vertices at approximately (1500,300), (1400,700) and (900,300) (all values in Hz). This region is similar to the 'nasal triangle' reported by Castelli (1989) and Pham (1995). Except in the case of /y/, the directions of the correlations of R1 and R2 (the long axes of the ellipses) tend to lie approximately parallel to the boundaries of the vowel triangle and the nasal triangle.

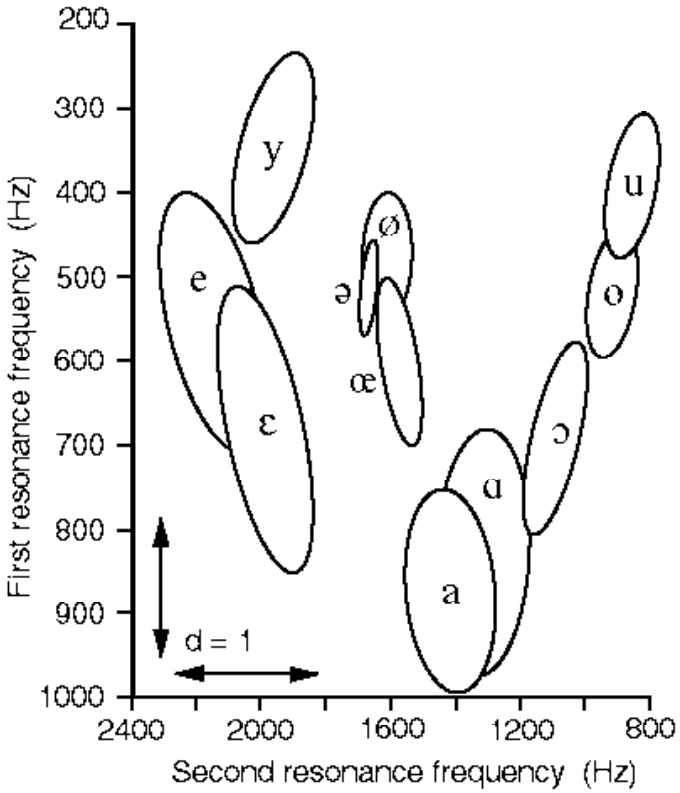

FIGURE 4. The vocal tract resonances (R2,R1) for seven adult female native speakers of French. The centre of each ellipse gives the mean values $(\bar{R2},\bar{R1})$. The major axis of the ellipse is parallel to the regression of R1 on R2, and the semi-axes are the standard deviations in that direction and in the perpendicular direction. The two lines at lower left show the unit of separation d defined by Equation 1. The relative sizes of the vertical and horizontal axes have been adjusted so that d has the same length in all directions. As is common practice in plotting formants, the direction of the axes is inverted, with the origin at the top right, in order to make the relative positions of the vowels the same as in conventional diagrams of the articulatory plane.

There appear to be few previous measurements of the vocal tract resonances for French, and not a very large number of published measurements of the formants. There are even fewer reports of measurements or estimations of formants for women's speech, and these are inevitably subject to greater inaccuracy than those of men's speech because the higher fundamental frequency means that harmonics are more widely spaced. Pham (1995) measured

[5] This failing was identified as an important limitation on the application of this technique. Since completion of this study, we have developed a new version of the hardware which includes, among other improvements, greater acoustic coupling between the source and the tract. This allows reliable measurement of vowels in all regions of the (R1,R2) plane (Epps, Smith and Wolfe, 1997).

the transfer function of the vocal tract for French vowels by stimulating the tract mechanically at the throat. This author plotted the two lowest resonances for five speakers, four male and one female (see Table 1). The relative positions of all vowels shown in Figure 4 are the same as those plotted by Pham. The values of R1 in this study are generally higher than those of Pham, which is not surprising because the speakers in this study were female. It is also possible that a group of Frenchwomen, largely Parisiennes, measured in Sydney might have a different accent from that of a sample of Frenchmen measured in Grenoble. Garnier-Rizet (1994) has reviewed measurements of formants for French vowels measured in seven previous studies with male speakers, including those reviewed by Durand (1985). These are compared in Table 1. The relative positions of the vowels are the same as those in Figure 4, but our measurements of R1 are again higher. The vocal tract in women is, on average, a little shorter than that in men, so one would expect R1 to be higher for women. There is the further difference that the data reviewed by Garnier-Rizet are for formants (here meaning the maxima in the speech spectrum) while those we report are for the resonances of the vocal tract. It is possible that the average formant frequency might not be the same as the resonance of the vocal tract because, for women speakers, the fundamental frequency is not negligible in comparison with the resonance frequency. It is possible, for instance, that estimates of F1 for female speakers might be skewed towards the fundamental frequency (F0) or the second or third harmonic (2 F0 or 3 F0), whichever were closer to R1 (see Figure 1).

		This study		Pham Thi Ngoc		Review of Garnier-Rizet	
		R1 Hz (s.d.)	R2 Hz (s.d.)	R1 Hz	R2 Hz	F1 (*s.d.(means)*)	F2 (*s.d.(means)*)
i	pie	-	-	351	2384	277 (*27*)	2321 (*157*)
e	thé	550 (90)	2150 (210)	391	2304	386 (*37*)	2188 (*171*)
ɛ	paix	680 (150)	1990 (170)	560	1924	524 (*45*)	1917 (*117*)
a	patte	870 (90)	1420 (180)	697	1323	756 (*76*)	1391 (*78*)
ɑ	pâte	830 (70)	1320 (190)	652	1198	777 (*87*)	1234 (*110*)
ɔ	pote	690 (100)	1090 (110)	566	1056	571 (*55*)	1138 (*101*)
o	pot	520 (40)	920 (100)	419	803	450 (*62*)	853 (*82*)
u	poux	390 (60)	860 (110)	364	708	279 (*21*)	811 (*51*)
y	pu	350 (70)	1960 (160)	363	1946	300 (*47*)	1863 (*91*)
ø	peu	470 (50)	1610 (90)	426	1545	453 (*64*)	1561 (*59*)
œ	peur	600 (90)	1580 (80)	576	1591	526 (*66*)	1445 (*100*)
ə	te	510 (50)	1670 (30)				

TABLE 1. The first two columns are the average frequencies, in Hz, of the first two resonances R1 and R2 of the vocal tract of the women native speakers used as targets in this study. The second two columns are R1 and R2 measured for a single woman native speaker by Pham (1995) using mechanical excitation near the glottis. The next two columns are the averages of formant frequencies found in the eight studies of male speakers reviewed by Garnier-Rizet (1994). Here the values in parentheses are the *standard deviation in the means* of those studies: the deviation amongst the values for different speakers would be rather greater.

Imitation of the targets.

Figure 5 shows the resonance frequencies of the vocal tracts of subjects imitating the six target vowels. In each plot, the head of the arrow represents the target. The tail is the mean result for the subjects and it is surrounded by an ellipse. As in Figure 4, the major axis of each ellipse is parallel to the linear regression of R1 on R2 for that set of points, and the semiaxes are the standard deviations in their directions. Thus long arrows mean poor matches and large ellipses mean much variability among subjects. In 5a (auditory feedback only) the target values (indicated by the heads of the arrows) are the resonances of the vocal tract of the native speaker whose recording was used for each vowel. When vocal tract feedback was used, the target

shown visually to the subjects was the average for each resonance over all native speakers, and this point is the target indicated by the heads of the arrows in 5b and c. (For the sound /ə/ in Figure 5c, the average of the imitations coincides with the target value, so no arrow is shown.)

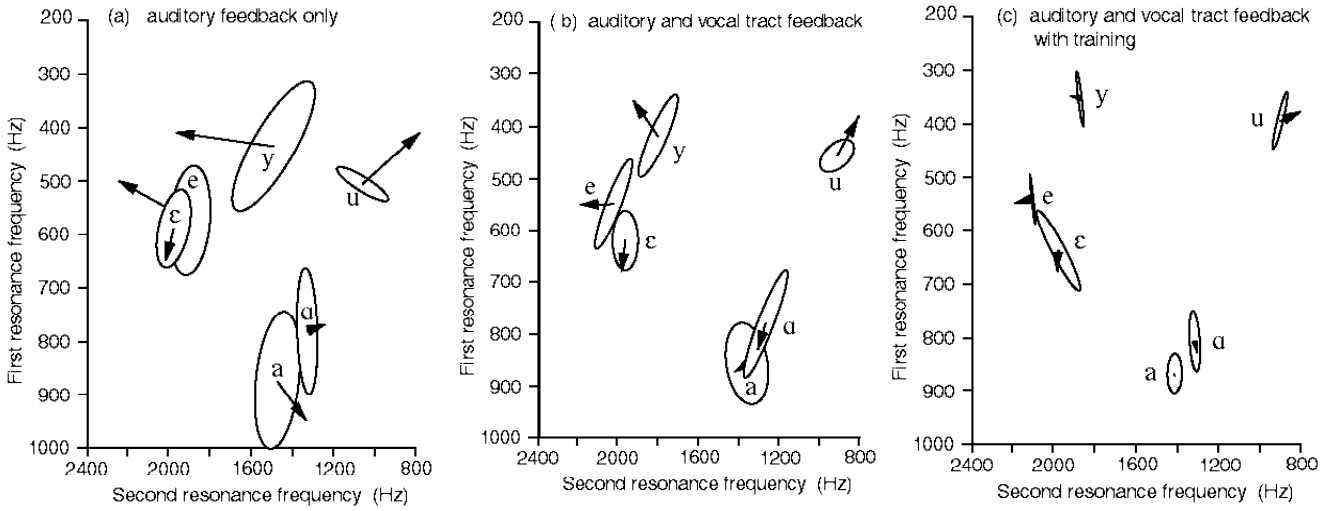

FIGURE 5. The resonance frequencies of the vocal tract (R2,R1) produced by imitation. The target values are indicated by the heads of the arrows; the tails of the arrows are the means of the values for the subjects performing the imitations. The major axis of each ellipse is parallel to the linear regression of R1 on R2 for that set of points, and the semiaxes are the standard deviations in their directions.

Figure 5 shows that the provision of vocal tract feedback improved the imitation, as measured by vocal tract response, and reduced the variability in the imitation. In the case of the subjects who had received one hour's training with vocal tract feedback in their first session (Figure 5c), the displacement from the target and the variability in the group were both rather smaller than the variation in (R2,R1) among the group of native speakers (Figure 4). For these subjects alone the spread of measured values is smaller than the distance separating /ə/ from /ɑ/ or /e/ from /ɛ/ in the (R2,R1) plane. The subjects who used only auditory feedback articulated /ʉ/ and /y/ with relatively little difference: their attempts are separated in (R2,R1) plane by a distance comparable with that separating /ə/ from /ɑ/. All subjects who used vocal tract feedback produced /ʉ/ separated from /y/ with a large difference in R2 (which suggests a large difference in place of articulation).

The difference in resonance frequencies can be quantified using d, the scaled distance in the (R2,R1) plane, as defined by Equation (1). Using this definition of d, two identical vocal tracts give d = 0, but for very different vowels d may be as high as 3. The differences in the average value of d among the three types of feedback shown in Figure 5 are significant at 99% (t test). We do not know what motor skills the subjects used to achieve the matches between the resonances of their own vocal tract and those of the target. Although R1 and R2 are determined primarily by jaw height and tongue position, they are also functions of other articulatory variables including mouth rounding and larynx height. Especially when both types of feedback were used, the subjects may have used a complex mixture of sensory cues and motor control to achieve good matches.

Although two identical vocal tracts give identical points in the (R2,R1) plane, it is not necessarily the case that vocal tracts with the same (R2,R1) produce the same sound. Higher resonances could differ, and the pitch and duration could vary. The ultimate test of pronunciation of a vowel is how well it is recognized by a native speaker. For that reason, we submitted recordings of the vowel sounds to a listening panel.

Results from listening panel.

The responses of the members of the listening panel to each of four different sets of sounds are shown in the confusion matrices (Figures 6a-6d). In each case, the twelve members of the panel decided which of the twelve non-nasalised vowels of French the recorded sound most closely resembled. The target vowels were the recordings of the vowels (one native speaker per vowel) which had been chosen as the targets for the subjects to imitate. For these the average identification score was 80%. This may seem at first to be low: after all most native speakers identify the words spoken by other native speakers with higher recognition rates than these. Identifying whole words, particularly when they occur in sentences, is usually rather easier than the identification of single phonemes because of the relatively high redundancy of spoken languages. There is the further complication of the somewhat awkward constraints imposed by the technical limitations of the experimental system employed: the subjects were asked to phonate a sustained vowel sound, to have their resonances measured, then to phonate again.

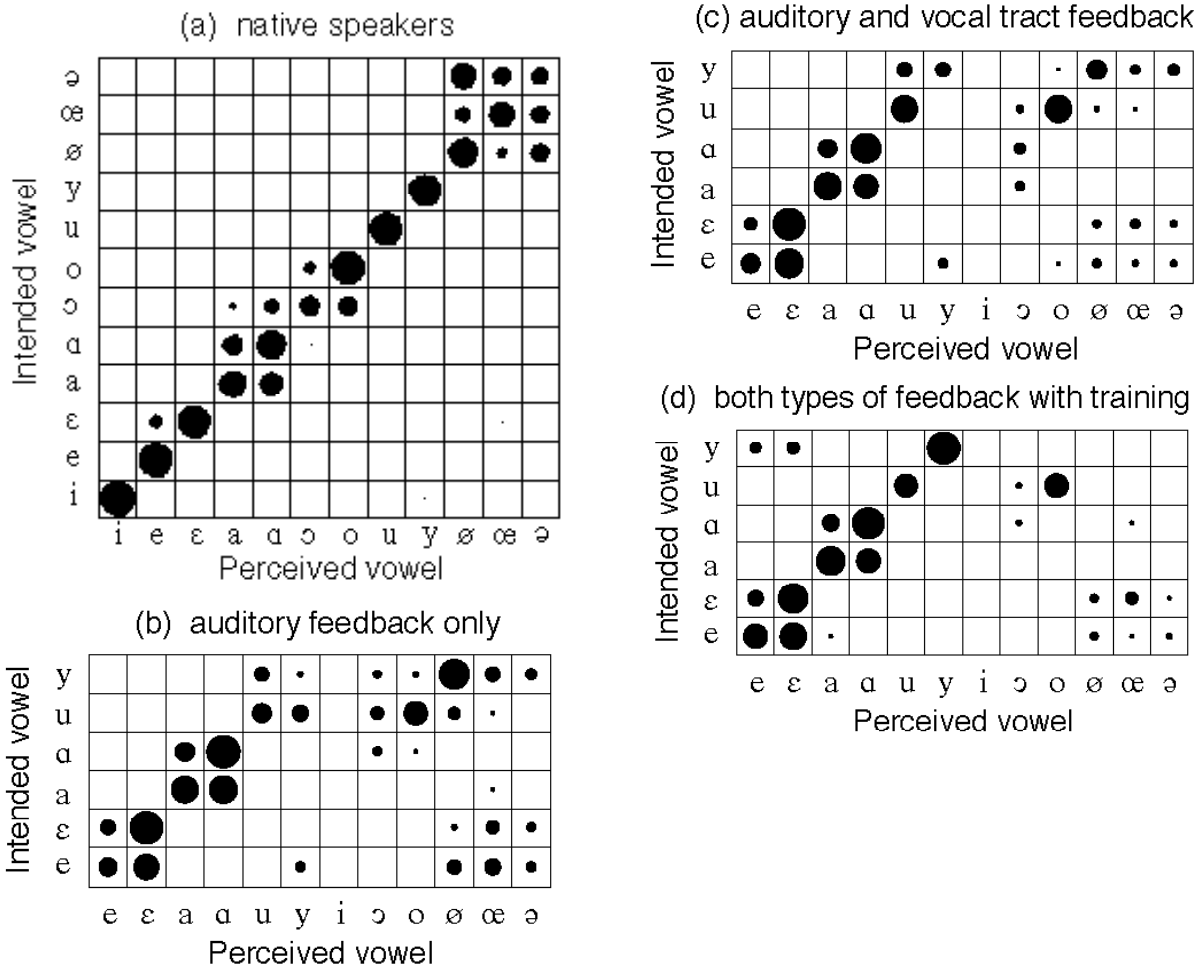

FIGURE 6. Confusion matrices. The rows correspond to the intended vowel, i.e. the vowel that each speaker was asked to produce (for native speakers) or was attempting to imitate (for subjects). The columns indicate the perceived vowel, i.e. the vowel that members of the listening panel judged it most resembled. Each entry is a circle with area proportional to the number of occurrences of that identification. 6a shows the results for the sounds made by the native speakers. 6b, c and d are for the sounds of the subjects who used different feedback treatments

The perceived vowel plane. All of the sounds heard by members of the listening panel were made by vocal tracts for which (R2,R1) was known (except for the vowel /i/). These sounds included the native speakers and the anglophone subjects from all treatments. The identification of the vowels most closely resembled by these sounds yields information about the

perceptual division or categorisation of the (R2,R1) plane by the members of the listeners' panel. Figure 7 displays the perceived vowel plane where each of the vowels is represented by an ellipse with centre at the mean value $(\overline{R2},\overline{R1})$ for all the sounds that were identified as that vowel (whatever the speaker intended them to be). As in Figure 4, the major axis of each ellipse is parallel to the linear regression of R1 on R2 for that vowel, and the semiaxes are the standard deviations in their directions. In most cases, the ellipses in Figure 7 (the categories in the perceived vowel plane) are larger than than those in Figure 4 (the distribution in the intended vowel plane), and this increased size is usually due to a larger distribution in the R2 direction. The exceptions are /ɛ/ and /ɑ/, where the sizes are smaller in the perceived vowel plane than in the intended vowel plane, and /e/ and /a/ which have similar sizes in the two planes. The central vowels /ø/, /ə/ and /œ/ have much larger distribution in the perceived plane than in the intended vowel plane. There are several possible reasons why there could be greater variation in the perception than in the production of vowels by native speakers. It could be because the formants (F2,F1) in the spoken sound cannot be determined as precisely as the resonances (R2,R1) of the vocal tract of the speaker. Further, it could be that, under the rather artificial conditions of our experiment, it is easier to produce vowels accurately than to identify them. Finally, it is likely that other information beyond the formants (F2,F1) is helpful in vowel identification. We return later to this point.

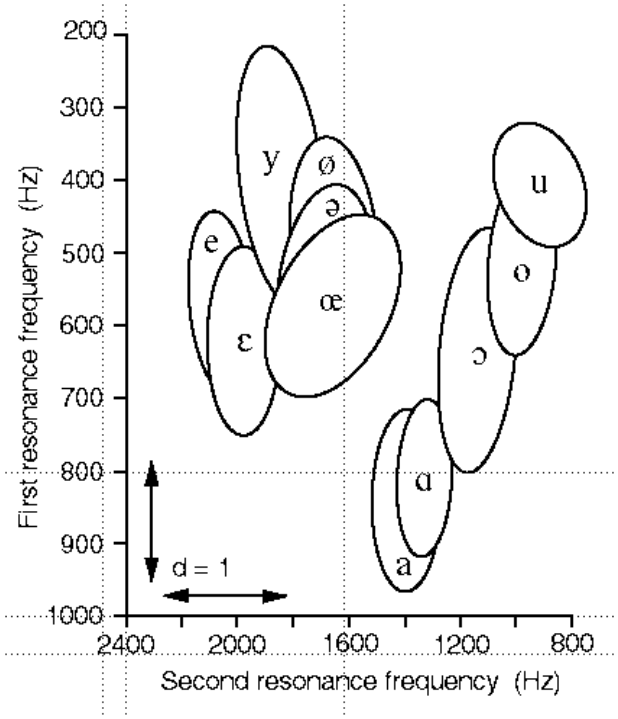

FIGURE 7.
Perceptual categorisation of sounds by native listeners. The centre of the ellipse for each vowel is at $(\overline{R2},\overline{R1})$, the mean values for all of the sounds that were identified as most closely resembling that vowel. The major axis of the ellipse is parallel to the regression of R1 on R2, and the semi-axes are the standard deviations in that direction and in the perpendicular direction. At lower left, two lines show the unit of separation d defined by Equation 1.

The listening panel results also allow the comparison of the separation d in the resonance plane and distinguishability by the members of the listening panel. The vowels of the native speakers plus the imitations by the anglophone subjects gave a set of points that sampled most of the vowel plane. Each decision made by a member of the panel associated a particular sound with a perceived vowel. This sound is positioned on the resonance plane by the (R2,R1) of the vocal tract that made it. The perceived vowel is best positioned on the resonance plane by the mean $(\overline{R2},\overline{R1})$ for that vowel as produced by all native speakers. Each listener's judgement of each vowel produced by each speaker is thus associated with a distance d between the (R2,R1)

and ($\overline{R2}$,$\overline{R1}$) just mentioned. Figure 8 presents a histogram of the fraction of the sounds identified as a function of d, their separation from the vowel that they were identified to be. For each identified vowel <V>, the sounds in an annulus around <V> were counted, and the number identified as <V> was expressed as a fraction ϕ to give a histogram of ϕ as a function of d for that vowel. For each value of d, the values of ϕ for each of the different identified vowels were averaged to give Figure 8.

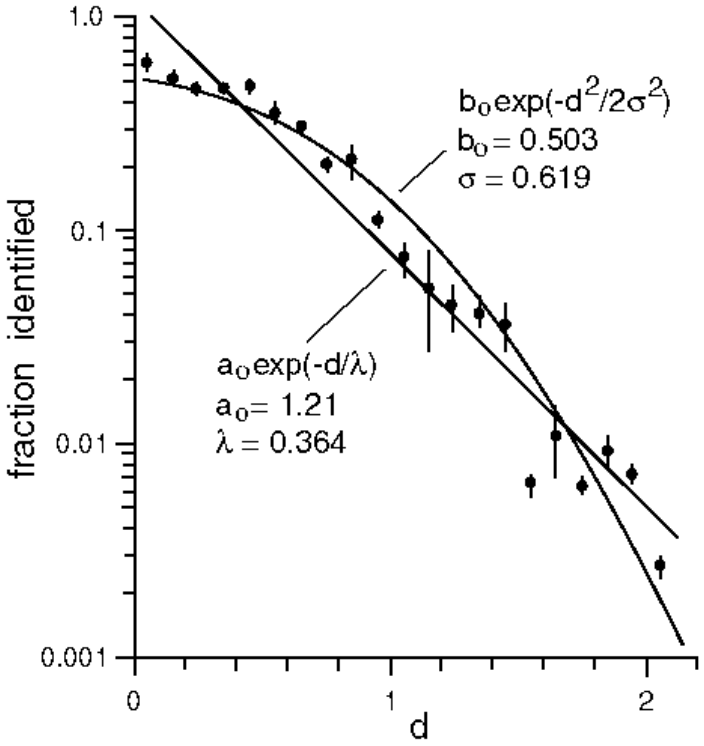

FIGURE 8. The fraction of identifications (particular sounds identified as vowels) as a function of the separation d between the particular sound (R2,R1) and the average for the perceived vowel ($\overline{R2}$,$\overline{R1}$). d is defined by Equation 1. The vertical axis is logarithmic and the error bars are ± one standard error. The straight line is the best fit to an exponential function (r = 0.98) and the curve is the best fit to a Gaussian function (r = 0.98).

The data in Figure 8 show that the chance of identifying a sound characterised by (R2,R1) as a vowel characterised by ($\overline{R2}$,$\overline{R1}$) decreases rapidly with the separation d between them. This result is not surprising. It is useful to fit a function to the data empirically because such a fit yields a characteristic value of vowel separation involved with distinguishability of vowels. Two simple functions are fitted in the Figure; both are moderately good fits to the data.

The values of λ or σ determined in this way are simple, dimensionless measures of the distinguishability of vowels in terms of their differences in R2 and R1, and they may be a characteristic of a particular language or dialect. It is not surprising that these values are of the same order as the smallest of the standard deviations in the R1 and R2 of individual vowels and of the same order as the smallest separations between adjacent vowels as produced by native speakers (Figure 4). One might expect λ and σ to be larger in a language with fewer vowels.

Confusion matrices. Figure 6a shows that the target vowel /ɛ/ was incorrectly identified by 25% of the listening panel of native speakers, who identified it as /e/. /a/ was incorrectly identified by 50% of the panel, who identified it as /ɑ/. We deliberately included /a/ and /ɑ/ because they had small separation d in the resonance plane (see Figure 4) and because they were regarded as very similar by native speakers. We included /ɛ/ and /e/ for the same reasons. There was no confusion between the vowels /ʊ/ and /y/ which have a large separation d and are recognised as being quite dissimilar by native speakers.

Figures 6b, c and d show confusion matrices for the monolingual anglophone subjects using the different combinations of feedback technique. There is considerable variability among the different subjects. The improvement provided by the vocal tract feedback however was not

uniform: it was greatest for the vowels /ʊ/ and /y/. The lack of significant improvement for the vowels /ɛ/ and /a/ is not surprising, as these vowels were poorly identified even when spoken by the native speakers whose recordings and vocal tract information were used as targets.

In the resonance plane (Figure 4) the target vowels /ɛ/ and /e/ are close together ($d = 0.29 < \lambda$), as are the target vowels /a/ and /ɑ/ ($d = 0.70$). These separations are of the same order as the variation among values of (R2,R1) measured in different native speakers for the same vowel (see Figure 4). This, along with unsolicited comments from the native speakers themselves, leads us to believe that information other than the resonances or formants is important in their distinction. This information appears to include duration in the case of the /a/ and /ɑ/, and possibly variation in fundamental frequency in the case of /ɛ/ and /e/. The vocal tract feedback we supplied gave no extra information on these parameters and so little improvement would be expected from its use. (F0 is displayed in a version of the apparatus developed since this study (Epps, Smith and Wolfe, 1997).)

Anglophones have difficulty distinguishing /ʊ/ and /y/ both in recognition and in production. For the subjects using auditory feedback, /y/ was poorly produced, and the majority of the listening panel identified it as /ʊ/ or /ø/ or another vowel. For the subjects trained with vocal tract feedback, however, the scores were better and the listening panel never identified an imitation of /ʊ/ as /y/, nor *vice versa*. In the case of these vowels, there is a substantial separation in the resonance plane ($d = 2.2 >> \lambda$) and so it is not surprising that the feedback relating to position in that plane gives the greatest improvement. Nevertheless, the performance of the group using vocal tract feedback in this case is a result which, if regularly reproduced, would be appreciated by those teaching French to anglophones. Averaging over all vowels studied, the values of recognition scores by the panel (standard deviations in parentheses) were: target vowels 80 (15)%, auditory feedback only 38 (24)%, both types of feedback 46 (19)%, and both types of feedback with training 63 (13)%. The recognition scores obtained by the group which had had a one hour session of training using vocal tract feedback were significantly better than the other imitations, at 95% confidence.

The significant difference provided by training with the vocal tract feedback is worth comment. In a previous study, in which one group of anglophone subjects imitated English vowels using vocal tract feedback without hearing the target vowels, we also observed significant improvement from first to second session (Dowd, Smith and Wolfe 1996). The vocal tract feedback method is new and requires an eye-mouth coordination which is quite unfamiliar. The subjects using this feedback had a total exposure to it of less than two hours by the end of their second session and might be expected to continue to improve with further use. The auditory feedback method, on the other hand, is a major component of the method used by most children to learn to speak their native language, and the subjects have had regular though informal exposure to auditory feedback ever since infancy. Further improvement might be expected from technological improvements which would allow subjects to phonate normally during measurements and which measured F0. We are currently developing such a system (Epps, Smith and Wolfe, 1997).

CONCLUSIONS

Direct measurements of the first two resonant frequencies of the vocal tract for eleven French vowels yield values similar to the formant frequencies measured from spectrograms by other researchers (reviewed by Garnier-Rizet, 1994), and their relative positions in the (R2,R1) plane are similar to those of the same vowels in the (F2,F1) formant plane. The confusion and correct identification of vowels are strongly related to their separation in the (R2,R1) plane. A sample of native speakers of English significantly improved their articulation and recognizability when pronouncing a set of these vowels using the real time measurement of the resonances of their own vocal tracts as visual feedback.

Acknowledgements

We acknowledge support from the the Australian Research Council. We also thank our volunteer subjects for their time and cooperation, Rosalind Epps and Xavier Boutillon for comments on the manuscript, and the journal's editor and reviewers who suggested several improvements.

Patents

The techniques described here are subject to provisional patent applications.

Appendix A

The impedance measurements were made using a spectrometer described in detail elsewhere (Wolfe et al 1994, 1995). Briefly, a waveform is synthesized as the sum of several hundred sine waves. It is amplified and input to a sealed loudspeaker (Figure 2) which is matched via an exponential horn to a high value acoustic resistor (R_{out} = 33 MΩ = 33 MPa.s.m^{-3}). This acts as a nearly ideal source of acoustic current. In the calibration procedure, the acoustic current is input to a reference load whose acoustic impedance is $Z_{ref}(f)$, where f is the frequency. The acoustic pressure is measured with a microphone and the pressure spectrum calculated. The measured spectrum $p_o(f)$ includes the frequency dependence of the amplifier, speaker, horn, acoustic resistor, microphone and reference load. A new waveform having Fourier components proportional to the reciprocal of those of $p_o(f)$ is then synthesized and input to the amplifier, speaker, horn, acoustic resistor, microphone and reference load. For a perfectly linear system, this would be expected to produce a completely flat output spectrum measured at the reference load. In this application, there are some non-linearities in the loudspeaker and so the iteration may be conducted three or more times in order to produce a frequency-independent pressure spectrum (a 'flat' spectrum) in the reference load $p_{ref}(f)$. The signal thus calibrated is then input to the spectrometer when it is coupled to the experimental load. The experimental load (impedance Z_{exp}) has a much lower value of impedance than the output resistance ($|Z_{exp}| << R_{out}$). While this condition is satisfied, the spectrometer output functions as a nearly ideal current source: i.e. the acoustic current with spectrum u(f) is independent of Z_{exp}. The measured pressure spectrum $p_{exp}(f)$ is therefore

$$p_{exp}(f) = u(f)\, Z_{exp} = \frac{p_{ref}(f)}{Z_{ref}(f)}\, Z_{exp}(f) = \text{constant}\, \frac{Z_{exp}(f)}{Z_{ref}(f)} \qquad (2).$$

The amplitude spectrum of $p_{exp}(f)$ from the microphone is calculated with a digital signal processor and displayed in real time.

In previous studies with this apparatus, we have usually used a low value resistive load as the reference (i.e. $Z_{ref}(f)$ has negligible imaginary component and is independent of frequency) so that the measured spectrum $p_{exp}(f)$ gives immediately $Z_{exp}(f)$. In this study, the reference load was chosen to be that of the laboratory field, measured at the lower lip of a subject with the mouth closed (Figure 2). This has two advantages for the application described here. First, the ratio of impedances then shows clearly the effect of opening the mouth and thus putting the vocal tract in parallel with the external field. Second, the acoustic impedance in a diverging wave increases with frequency and is small at low frequencies. Using a frequency-independent acoustic current leads to poor signal to noise ratios at low frequencies. By using a diverging wave as the reference, the calibration procedure gives larger acoustic currents at low frequencies and thus improves the signal to noise ratio in this range.

We restricted the frequency range of the acoustic current to 200 to 2500 Hz, with components every 20 Hz. The lower limit was below the lowest expected value of R1. The higher limit often restricted us to the two resonances R1 and R2 in which we were most interested. When higher resonances fell within this range their effects on the impedance were usually much smaller in magnitude than those of R2, and we did not draw the attention of the subjects to them.

References

CASTELLI, E. (1989). *Caractérisation acoustique des voyelles nasales du français. Mesures, modélisation et simulation temporelle*. Doctoral thesis, INPG, Grenoble, cited by Pham Thi Ngoc (1995).

CLARK, J. and YALLOP, C. (1990). *An Introduction to Phonetics and Phonology*, Basil Blackwell Ltd, Oxford.

DJERADI, A., GUERIN, B., BADIN, P and PERRIER, P. (1991). Measurement of the acoustic transfer function of the vocal tract: a fast and accurate method. J. Phonetics, **19**, 387-395.

DOWD, A., SMITH, J. and WOLFE, J. (1996). Real time, non-invasive measurements of vocal tract resonances: application to speech training. Acoustics Australia, **24**, 53-60.

DURAND, P. (1985). *Variabilité acoustique et invariance en français, consonnes occlusives et voyelles*. CNRS, Collection *Sons et Parole*, Paris.

EPPS, J., SMITH, J.R. and WOLFE, J. (1997). A novel instrument to measure acoustic resonances of the vocal tract during speech. Measurement Science and Technology, **8**, 1112-1121.

FANT, G. (1973). *Speech Sounds and Features*. MIT, Cambridge, Mass.

GARNIER-RIZET M. (1994). *Élaboration d'un module de règles phonético-acoustiques pour un système de synthèse à partir du texte pour le français*. Doctoral thesis, Université de Paris III.

HÖGBERT, J. (1995). From sagittal distance to area function and male to female scaling of the vocal tract. Quarterly Progress and Status Report, Dept. of Speech, Music and Hearing, KTH , 1995 No. 4 pp 11-53.

LANDERCY, A. and RENARD, R. (1977). *Éléments de Phonétique*. Didier, Bruxelles.

LIÉNARD, J-S. (1977). *Les processus de la communication parlée. Introduction à l'analyse et la synthèse de la parole*. Masson, Paris.

LINDBLOM, B.E.F. and SUNDBERG, J.E.F. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. JASA **50**, 1166-1179.

MAKHOUL, J. (1975). 'Linear prediction: a tutorial review', Proc. IEEE, **63**, 561-579.

PHAM THI NGOC, Y. (1995). *Caractérisation acoustique du conduit vocal: fonctions de transfert acoustiques et sources de bruit*. Doctoral thesis, Institut National Polytechnique de Grenoble.

PHAM THI NGOC, Y. and BADIN, P. (1994). Vocal tract acoustic transfer function measurements: further developments and applications. J. de Physique IV, **C5**, 549-552.

REY, A. and REY-DEBOVE, J. (eds) (1985). *Le Petit Robert: Dictionnaire Alphabétique et Analogique de la Langue Française*. Dictionnaires le Robert, Paris.

SUNDBERG, J. (1987). *The Science of the Singing Voice*, Northern Illinois Univ. Press., De Kalb, Ill.

WOLFE, J., SMITH, J., BRIELBECK, G. and STOCKER, F. (1994). Real time measurement of acoustic transfer functions and acoustic impedance spectra. *Australian Acoustical Society Conference*, Canberra, pp. 66-72.

WOLFE, J., SMITH, J., BRIELBECK, G. and STOCKER, F. (1995) A system for real time measurement of acoustic transfer functions. Acoustics Australia, **23**, 19-20.